

Sustainable Indoor Trajectory Modeling through Participatory Sensing and WiFi Signal Fusion

Konstantinos Vogklis
Department of Tourism
Ionian University
Corfu, Greece
voglis@ionio.gr

Ioannis Nasios
Nodalpoint Systems
Athens, Greece
ioannis.nasios@nodalpoint.com

Anastasia Vayona
Circular Economy Foundation
Brussels, Belgium
anastasia.vayona@circular.foundation

Vasilis Katos
School of Computing and Engineering
Bournemouth University
Bournemouth, UK
vkatos@bournemouth.ac.uk

Abstract—In this paper a general machine learning approach for offering indoor location awareness without the need to invest in additional and specialised hardware is presented. We explore use cases where visitors equipped with their smart phones would interact with the available WiFi infrastructure to estimate their location, since the indoor requirement poses a limitation to standard GPS solutions. The need for human participation, also referred to as human capital, is highlighted as a critical factor to the success of the approach, in order to compensate for the omission of specialized indoor location estimation equipment. Furthermore, the proposed framework enables knowledge discovery from fragmented indoor mobility trajectories by fusing sensor signals, aligning with recent advances in multi-sensor trajectory analytics. The results have shown that the proposed approach achieves accuracy of less than 2 m and in the case where a substantial number of BSSIDs are dropped, the fusion has the ability to maintain the accuracy of the approach.

Index Terms—WiFi signal fusion, indoor location estimation, participatory sensing, mobile crowd sensing

I. INTRODUCTION

Sustainability has become a central objective in the development of smart cities [1], [2], yet the transition to smart infrastructures often requires significant investment in hardware, resources, and human capital. This may contradict sustainability goals, as highlighted by the Jevons' paradox [3]. The circular economy paradigm addresses this challenge by promoting the efficient use and repurposing of existing resources through practices such as reuse, rethink, and redesign [4], [5]. In this context, data-driven circular economy leverages information and communication technologies (ICT) to create value while minimizing resource consumption [6].

In this paper, we demonstrate how existing assets, smartphones, WiFi infrastructure, and citizen participation, can be repurposed to provide accurate indoor location services without requiring additional hardware. We present an end-to-end framework covering data collection through mobile crowd sensing (MCS), preprocessing, and machine learning-based modeling for indoor positioning.

Our approach highlights the role of citizen engagement as a substitute for infrastructure investment, showing how social capital can enhance system performance while maintaining sustainability principles. Increased participation leads to improved localization accuracy without increasing hardware demands, enabling a scalable and resource-efficient solution.

The proposed framework lies at the intersection of MCS and machine learning (ML). MCS enables the collection and fusion of data from user devices, while ML supports the extraction of meaningful patterns for location estimation [7], [8]. This combination allows the development of accurate and adaptive indoor positioning services, particularly in environments where GPS is unreliable.

From a technical perspective, we focus on WiFi-based indoor localization using signal strength fusion. The proposed method achieves high accuracy while remaining resilient to missing or unstable signals, reducing the need for frequent retraining. Empirical evaluation demonstrates that accurate indoor positioning can be achieved using only existing infrastructure, supporting sustainable and scalable smart city applications.

II. LOCATION BASED SERVICES

Location-based services (LBS) are widely used in mobile applications, enabling context-aware functions such as navigation, marketing, and social networking [9], [10]. Their effectiveness depends on accurate positioning, which remains challenging in indoor environments where GPS is unreliable.

To address this limitation, alternative technologies such as WiFi, RFID, and ultra-wideband have been developed [10], [11], with WiFi-based approaches being the most widely adopted due to existing infrastructure.

Indoor localization using WiFi typically relies on fingerprinting techniques that map signal strengths (RSS) to known locations. These approaches can be categorized into probabilistic, deterministic, and ML methods. ML models, including

KNN and decision trees, have demonstrated strong performance in capturing complex signal patterns and improving localization accuracy [12].

Recent large-scale studies have shown that combining WiFi signals with additional sensor data can achieve positioning accuracy between 1.5 m and 3 m [13]. Building on these approaches, recent research has increasingly focused on improving robustness, scalability, and resilience to environmental and device variability without relying on additional infrastructure. In this context, hybrid methods that combine WiFi fingerprinting with motion-based techniques have demonstrated enhanced localization performance [14], while ensemble-based approaches have further improved adaptability and generalization across heterogeneous deployment conditions [15].

Recent advances in indoor location-based services highlight the importance of data quality, learning-based inference, and privacy-aware system design. In particular, WiFi fingerprinting remains a dominant paradigm due to its scalability and compatibility with existing infrastructure, yet its performance is highly dependent on the quality of the radio map. Data cleansing techniques have been shown to improve positioning accuracy and efficiency by removing low-correlation or noisy fingerprints, leading to measurable gains in both 2D/3D localization error and computational performance [16]. At the same time, the integration of deep learning models with enhanced signal features, such as WiFi fine time measurements (FTM), enables sub-meter localization accuracy by learning complex spatial representations from heterogeneous signal inputs [17]. Beyond accuracy, recent research also emphasizes the need for privacy-preserving indoor positioning systems, where user participation and localization processes are designed to minimize exposure of sensitive information while maintaining service utility. In this broader context, indoor localization can be viewed as a mobile data management (MDM) problem involving noisy, crowdsourced, and privacy-sensitive spatio-temporal data streams, aligning closely with the core themes of the MDM research community.

Despite these advances, most existing approaches primarily focus on improving accuracy, with limited consideration of sustainability and citizen participation. This work addresses this gap by exploring WiFi-based localization through data fusion and MCS, without requiring additional hardware deployment.

III. EXPERIMENTAL SETUP

To address the technical challenges and to illustrate the principles of MCS, an experiment was conducted at indoor premises, employing the existing WiFi infrastructure of a building situated at Bournemouth University in the UK. This venue was selected as it provided ease of access to the ceiling that had already a grid of lights with an ideal to the experiment distances, as well as adequate spatial complexity (straight and looping corridors, and open spaces) and size.

The first task was to establish the ground truth representation. This was achieved by annotating the ground floor map shown in Figure 1 with a grid of labeled points. Each labeled

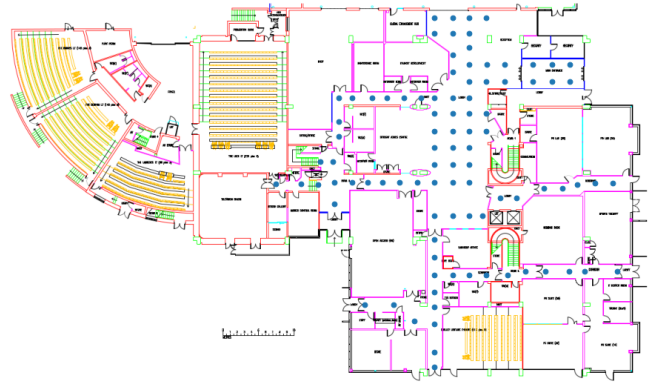


Fig. 1. Ground Floor Map with Label Points

point corresponds to a set of indoor coordinates in meters where (0,0) was assigned to the bottom right corner of the building. In Figure 1, we show approximately 100 ground truth points that form a grid of approximately 2.5 meters wide. In this way, we create a set of N_{max} checkpoints $\mathcal{X} = \{(x_i, y_i), \text{ for } i = 1, \dots, N_{max}\}$, which will represent known coordinates within the building in meters.

In order to create the initial dataset, having defined the initial set of checkpoints, an Android data acquisition application was created and shared with the participants. The users were instructed to hold the mobile device under the label and tap on the corresponding number displayed on their screen. The data acquisition application included a connected graph of all labels in order to provide easy access to the users by presenting only the adjacent labels (nodes). When the user clicks on a label, the application records the specific label ID¹ with the current timestamp and the screen refreshes with the labels adjacent to the selected point.

The above data collection protocol would allow a user to seamlessly navigate within the building and create predefined or ad-hoc paths.

The raw data from the application are captured and stored in a series of text log files, where each file represents a unique recorded path. The user can decide when to begin or end a specific path. In addition to user input at the labeled points, the application records a list of surrounding WiFi BSSIDs² along with its signal strength in dB every 2 seconds. In addition, sensor data from the mobile device, including gyroscope, accelerometer, and magnetometer readings are recorded every 50 milliseconds. These are included in the published dataset to enable current and future research on dead reckoning [18], [19].

IV. MACHINE LEARNING MODELING

The aim of the proposed approach is to create/train a regression model that would take a list of nearby WiFi BSSIDs

¹Also referred as checkpoint

²BSSID stands for Basic Service Set Identifier. This is the MAC physical address of the access point or wireless router that is used to connect to the WiFi.

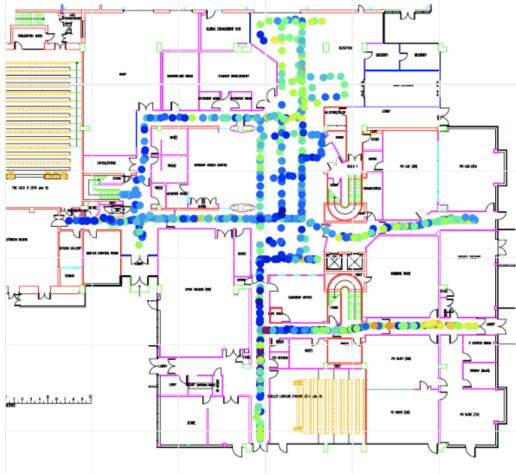


Fig. 2. Ground Floor Map with a sample path

along with the corresponding signal strength levels as input and will approximate the internal coordinates (x_i, y_i) at the predefined checkpoints $\mathcal{X} = \{(x_i, y_i), \text{ for } i = 1, \dots, N_{max}\}$. This is a two-step process: a) first, the data are prepared and expressed in an appropriate tabular/numeric format and b) a suitable parametric regression model is estimated through training.

To suit the participatory concepts of the proposed approach, the training procedure must be able to incorporate new data when they are available (online training) and in order to continuously improve - or perhaps avoid any detrimental effects to - the predictions.

A. Data pre-processing

During the citizen participation phase, the routes are captured using a mobile device and stored in the form of a text log file, one for each distinct route. Routes that contain more than 4 checkpoints are only saved and considered in order to avoid really short and probably erroneous paths. Then all available WiFi present signals across all paths are aggregated to calculate the frequency of appearance of each and every BSSID. To keep one BSSID as input element for the modeling phase, it is required that this BSSID appear at least 50 times within the dataset. This makes the approach more robust as it allows the preprocessor to discard remote WiFi antennas from other buildings that are visible in a limited area. Furthermore, WiFi SSIDs containing smartphone model keywords such as *Android*, *Galaxy*, *iPhone*, *HUAWEI*, *Pixel*, *Nokia*, *Honor* were also discarded, as they correspond to mobile hotspots. These SSIDs are not fit for purpose and are highly volatile, as they are expected to change position or disappear, causing deterioration to the predictions.

A crucial training parameter is the number of unique BSSIDs collected from all paths. In the experimental testbed $N_{bssid} = 312$ unique BSSIDs were recorded. Each row of the dataset corresponds to a checkpoint and contains the (x, y) position in meters to be used as a target feature. The input

vector is N_{bssid} -dimensional and contains the signal strength of visible WiFi around the specific checkpoint or a value representing an unknown set to -999 otherwise.

The initial dataset consists of 14,373 rows that correspond to 423 unique paths. Each data row contains a timestamp indicating when this sample was recorded and the path name to which this sample belongs. This yielded 423 unique paths with an average of approximately 34 rows - timestamps per path.

Some of the WiFis were detectable only 46 times (0.3%) and some others 12,823 times (89.2%) with a median of 1,300 (9%).

B. Modeling experimental setup

The initial investigation involved three diverse regression models: (a) gradient boosting trees (GBM) referred as LightGBM [20] (b) recurrent neural network (RNN) using Keras - tensorflow implementation [21] and (c) k-nearest neighbors (KNN) [22] which is an unsupervised approach. All three models, despite adopting considerably different approaches, have been accepted to be suitable for this task, as evidenced from the best solutions in a relevant ML competition [13].

The evaluation protocol is *5-fold cross validation* and folds are *split by path* so that a complete path belongs to a single fold. For comparative analysis, the split was kept the same for all experiments. The optimized metric, is the *mean position error* that is, the Euclidean distance in meters:

$$MPE = \sum_{i=1}^{N_{max}} \sqrt{(x_i - \tilde{x}_i)^2 + (y_i - \tilde{y}_i)^2}, \quad (1)$$

where:

N_{max} is the number of rows in the dataset

(x_i, y_i) are the ground truth locations for a given test row

$(\tilde{x}_i, \tilde{y}_i)$ are the predicted locations for a given test row

C. Experimental evaluation

The experimental investigation initially involves a comparison of the multiplicity of errors (MPE) of each of the selected regression models. The GBM models were configured with low learning rate, deep tree structures, and feature/bagging subsampling for improved generalization. For the implementation of the recurrent neural network, LSTM [23], [24] layers were used with the architecture combining Dense preprocessing layers with stacked LSTM units and regularization (Batch Normalization, Dropout). The simple KNN model consists of a single parameter set to $k = 45$.

As depicted in Table I, GBM performs better than the other two models, followed by RNN. Furthermore, merging the output of these models in a 2 : 1 ratio reduces the mean position error. Lastly, GBM exhibits the smallest standard deviation of MPE across 5-folds, which is a desirable performance indicator.

In the remaining subsections, a set of experiments is presented to evaluate how different conditions affect the MPE

TABLE I
MEAN POSITION ERROR OF DIFFERENT MODELS

Model	Mean MPE	Std MPE
GBM	1.95	1.53
RNN	2.08	1.69
KNN	2.40	1.76
best ensemble	1.92	1.53

metric. The analysis focuses on two key data-driven dimensions: data paths and infrastructure. Regarding data paths, we examine how the amount of collected path data influences prediction accuracy, as well as the level of effort required from citizens to achieve a sufficiently low error. In terms of infrastructure, we investigate the role of WiFi/BSSID data, including the minimum number of access points required for effective model training, the impact of grid size on accuracy, and the robustness of the model under conditions where BSSIDs are partially unavailable during prediction.

1) *Impact of the size of the data set on overall accuracy:* A sensitivity analysis was conducted by repeatedly training the models (five runs) on progressively smaller random subsets (75%, 50%, 25%) of the total 423 paths to assess the impact of data volume on positioning accuracy. Results, shown in Table II, indicate that increasing the number of paths improves performance; however, gains diminish beyond a certain point, with reductions in mean positioning error (MPE) reaching only centimeter-level improvements. This suggests the existence of a saturation threshold, beyond which additional data (i.e., user participation) yields limited accuracy benefits.

TABLE II
PERFORMANCE OF DIFFERENT MODELS WITH INCREASING NUMBER OF PARTICIPANTS (PATHS)

Model	Paths	Repeats	Mean MPE
GBM	106	50	2.21
RNN	106	50	2.40
KNN	106	50	2.62
GBM	211	40	2.06
RNN	211	40	2.25
KNN	211	40	2.48
GBM	317	30	1.99
RNN	317	30	2.15
KNN	317	30	2.43
GBM	423	1	1.95
RNN	423	1	2.08
KNN	423	1	2.40

2) *Node fusion: Impact of number of different BSSIDs on accuracy:* The experiment involved the collection of 312 different WiFi BSSIDs. By removing a fixed percentage of BSSIDs and simulating conditions where some BSSID signals never existed, we can further measure the performance of the models.

We used 75%, 50% and 25% of the 312 unique WiFi BSSIDs (234, 156 and 78 WiFi). As some WiFi may be more

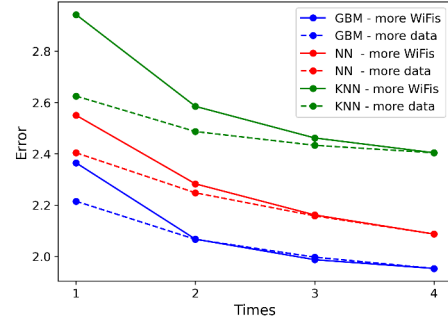


Fig. 3. Models performance for increasing number of participants (paths) MHPWS: Comparison of models performance on increased WiFi or data

important than others, we ran our experiments multiple times for every model and each time, a different subset of WiFi was selected in random. The fewer the available WiFi, the more experiment repetitions was performed in order to obtain more accurate results as the standard deviation increased. The results are shown in Table III.

As expected, fusing more WiFi signals leads to more accurate location estimation. The results could have been improved if there had been even more WiFi in place. However, inspecting the trend between the available WiFi and the score, it is argued that the actual performance is too far from saturation.

Figure 3 shows the performance increase when introducing WiFi points and is compared with that of more data (paths). The trade-off is apparent and shows the preference of having more WiFi over path data.

TABLE III
PERFORMANCE OF DIFFERENT MODELS WITH LESS WiFi ANTENNAS INSTALLED

Model	WiFi	Repeats	Mean Error	Std
GBM	234	30	1.98	0.01
RNN	234	30	2.16	0.02
KNN	234	30	2.46	0.02
GBM	156	40	2.06	0.03
RNN	156	40	2.28	0.03
KNN	156	40	2.58	0.04
GBM	78	50	2.36	0.08
RNN	78	50	2.55	0.06
KNN	78	50	2.94	0.10

3) *Measuring the resilience of fusing:* The models were trained with 312 WiFi BSSIDs. To assess robustness, we simulate scenarios where BSSIDs become unavailable due to faults, replacements, or larger disruptions (e.g., fire affecting the operational map).

Performance was evaluated by progressively removing BSSIDs at random. Results in Figure 4 demonstrate strong resilience: all models maintain acceptable accuracy up to 50 missing BSSIDs (about 16%), while neural networks and nearest neighbors remain stable up to 150 missing BSSIDs.

This indicates that WiFi signal fusion can tolerate moderate signal loss without significant degradation.

Fusing model predictions further improves robustness, consistently outperforming individual models across all scenarios. The advantage becomes more pronounced beyond 250 missing BSSIDs (approximately 80%).

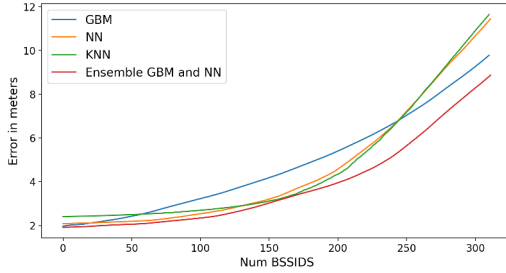


Fig. 4. Model Performance plot over various missing WiFis

Beyond this point, GBM slightly outperforms the other models, although all errors increase above 6 m. While such accuracy may suffice for coarse applications (e.g., locating vulnerable users), it is inadequate for more sensitive use cases such as social distancing or occupancy monitoring.

4) *Impact of target grid on overall accuracy:* The target grid spacing on the testbed is approximately 2.32 m. By removing every second intermediate node, the data set is split into two new data sets with different, no overlapping target nodes. As a result, each data set defines a grid with a 3.91 m spacing. The results presented in this subsection are the average of two training pipelines, one for each sub-dataset.

TABLE IV
PERFORMANCE OF DIFFERENT MODELS WITH LARGER TARGET SPACING

Model	same split	other split	average
GBM	1.89	2.17	2.03
RNN	1.92	2.46	2.19
KNN	2.37	2.56	2.46
GBM + RNN	1.82	2.18	2.00
2*GBM + RNN	1.83	2.14	1.99

After completion of training with half the original data and with a more sparse target grid, the performance of each model and the ensemble model decreased as expected. As shown in Table IV, the first column (same split) shows the model performance for the data belonging to the same split and the second column (other split) shows the performance for the data from the complementary split, with the training target nodes missing. The third column is the average of these two columns representing the true model performance in this setup. Figure 5 shows the two sets of targets (blue and red dots) as well as individual predictions (of the same split) of magnitude, where the larger the dot, the larger the prediction error.

V. DISCUSSION

Gathering all necessary data by randomly walking under the infrastructure, added diversity to the final dataset. Subscribing

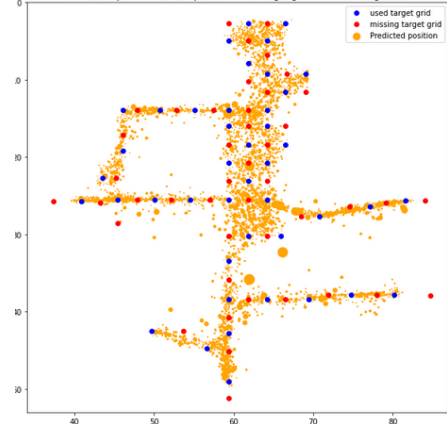


Fig. 5. The effect of the two complementary grid sets on prediction (the radius of the orange dot is relative to the estimation error).

to the common position that technology alone is not sufficient in achieving sustainability, we eluded how the social capital of a city can be recruited in providing a primitive function on a technical and operational level by both respecting the circular economy (CE) paradigm and encouraging the hospitality sentiment of a city.

Fusing all detectable WiFi signals every 2 seconds will create a location path upon movement. Such a path can be seen in Figure 6 where the GBM's (x, y) predictions are represented as blue points and the ground truth as orange points. The macro structures (hallways and rooms) within the building are clearly visible. A path of a single participant (ground truth) is depicted with red connected points. It is visually evident that the GBM model manages to generalize and create meaningful predictions even between successive ground truth points. Some sparse outliers exist, but they can be easily filtered out by a smoothing process.

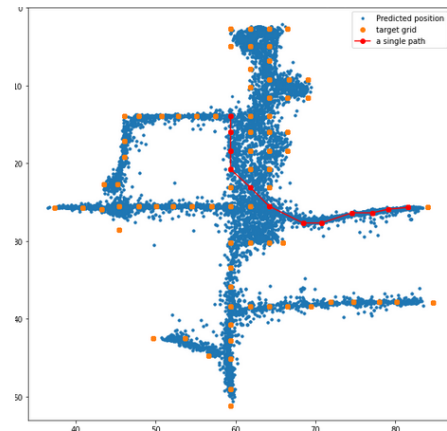


Fig. 6. GBM predictions over target grid

An example of the predictions of different models for one single path is shown in Figure 7. The same path is also marked, with a red line, in Figure 6. For this path, the mean

errors are: 1.25 m for gradient boosting machine, 1.52 m for Neural Network and 1.91 m for nearest neighbors model. At the left side of the plot, predictions for all models are further away from the true path, while at the right part of the plot, predictions are forming a path that matches with the true path. From this plot alone, it is easy to distinguish a hallway (right) from a larger room (left).

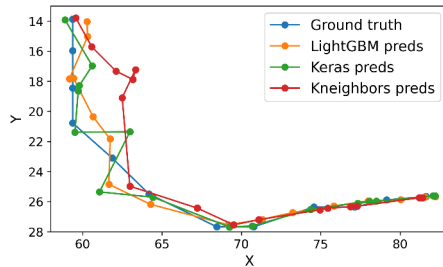


Fig. 7. Models Predictions of a selected path

VI. CONCLUSIONS AND FUTURE WORK

Indoor location tracking is a valuable primitive function for a number of use cases. It was demonstrated in this research that indoor location awareness while respecting sustainability principles is attainable by employing MCS and ML. Indoor location prediction with less than 2 m error was achieved by fusing data from available WiFi signals only. The three models were trained out of which both GBM and RNN gave best performance for the task. Furthermore, fusing GBM with RNN estimations not only increased the overall performance but also resulted in a significant improvement in resilience.

While the testbed consisted of the ground floor of a relatively large building with a diverse layout, future experiments could leverage the data annotation and training protocols proposed in this work to construct and explore more complex structures and paths (such as moving along the Z axis by taking a lift or a staircase), as well as scenarios such as crowd management, real-time social distancing, dynamic fire evacuation plans, including the support of vulnerable citizens.

The dataset used for this research is also made available. Apart from the WiFi signals, it also contains data from the internal sensors (accelerometer, gyroscope, rotation, magnetometer) that, when combined with the annotated grid points, they can be used to promote research on dead reckoning techniques, offering even higher levels of resilience to indoor location capabilities.

The proposed approach enables a wide range of applications, including privacy-aware emergency response (e.g., nearest doctor identification), energy-efficient building management through occupancy-aware analytics, and adaptive indoor space planning for scenarios such as social distancing.

REFERENCES

[1] M. A. Fadhel, A. M. Duhaim, A. Saihood, A. Sewify, M. N. Al-Hamadani, A. S. Albahri, L. Alzubaidi, A. Gupta, S. Mirjalili, and Y. Gu, "Comprehensive systematic review of information fusion methods in

smart cities and urban environments," *Information Fusion*, vol. 107, p. 102317, 2024.

[2] B. P. L. Lau, S. H. Marakkalage, Y. Zhou, N. U. Hassan, C. Yuen, M. Zhang, and U.-X. Tan, "A survey of data fusion in smart city applications," *Information Fusion*, vol. 52, pp. 357–374, 2019.

[3] J. M. Polimeni and R. I. Polimeni, "Jevons' paradox and the myth of technological liberation," *Ecological Complexity*, vol. 3, no. 4, pp. 344–353, 2006.

[4] G. Bressanelli, F. Adrodegari, M. Perona, and N. Saccani, "Exploring how usage-focused business models enable circular economy through digital technologies," *Sustainability*, vol. 10, no. 3, p. 639, 2018.

[5] J. Potting, M. Hekkert, E. Worrell, and A. Hanemaaijer, *Circular Economy: Measuring innovation in the product chain*, 01 2017.

[6] D. J. Langley, J. Van Doorn, I. C. Ng, S. Stieglitz, A. Lazovik, and A. Boonstra, "The internet of everything: Smart things and their impact on business models," *Journal of Business Research*, vol. 122, pp. 853–863, 2021.

[7] J. A. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava, "Participatory sensing," 2006.

[8] B. Guo, Z. Yu, X. Zhou, and D. Zhang, "From participatory sensing to mobile crowd sensing," in *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*, 2014, pp. 593–598.

[9] H. Huang, G. Gartner, J. M. Krisp, M. Raubal, and N. Van de Weghe, "Location based services: ongoing evolution and research agenda," *Journal of Location Based Services*, vol. 12, no. 2, pp. 63–93, 2018.

[10] D. Mallis, C. M. Angelopoulos, V. Katos, and K. Vogklis, "Isums: Indoor space usage monitoring system for sustainable built environment using lorawan," in *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2021, pp. 473–482.

[11] G. Shi and Y. Ming, "Survey of indoor positioning systems based on ultra-wideband (uwb) technology," in *Wireless communications, networking and applications*. Springer, 2016, pp. 1269–1278.

[12] A. Nessa, B. Adhikari, F. Hussain, and X. N. Fernando, "A survey of machine learning for indoor positioning," *IEEE Access*, vol. 8, pp. 214945–214965, 2020.

[13] Kaggle, "Indoor location navigation," <https://www.kaggle.com/competitions/indoor-location-navigation/>, 2021.

[14] B. Bonthu and S. Mohan, "Combining wi-fi fingerprinting and pedestrian dead reckoning to mitigate external factors for a sustainable indoor positioning system," *Sustainability*, vol. 15, no. 14, p. 10943, 2023.

[15] P. Roy, C. Chowdhury, D. Ghosh, and S. Bandyopadhyay, "Juindoor-loc: A ubiquitous framework for smartphone-based indoor localization subject to context and device heterogeneity," *Wireless Personal Communications*, vol. 106, pp. 739–762, 2019.

[16] D. Quezada-Gaibor, L. Klus, J. Torres-Sospedra, E. S. Lohan, J. Nurmi, C. Granell, and J. Huerta, "Data cleansing for indoor positioning wi-fi fingerprinting datasets," in *Proceedings of the IEEE International Conference on Mobile Data Management (MDM)*, 2022.

[17] P. Eberechukwu, H. Park, C. Laoudias, S. Horsmanheimo, and S. Kim, "Dnn-based indoor fingerprinting localization with wifi ftm," in *Proceedings of the IEEE International Conference on Mobile Data Management (MDM)*, 2022.

[18] M. M. del Horno, L. Orozco-Barbosa, and I. García-Varea, "A smartphone-based multimodal indoor tracking system," *Information Fusion*, vol. 76, pp. 36–45, 2021.

[19] S. Qiu, Z. Wang, H. Zhao, K. Qin, Z. Li, and H. Hu, "Inertial/magnetic sensors based pedestrian dead reckoning by means of multi-sensor fusion," *Information Fusion*, vol. 39, pp. 108–119, 2018.

[20] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017, pp. 3146–3154.

[21] F. Chollet *et al.*, *Deep learning with Python*. Manning New York, 2018, vol. 361.

[22] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[23] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[24] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.